The Report committee for Yutong Duan

certifies that this is the approved version of the following report:

Implementing the Multimodel Generalized Beta Estimator in Stata and Its Application

APPROVED BY

SUPERVISING COMMITTEE

Supervisor: _____

Paul T. von Hippel

Co-Supervisor: _____

Mingyuan Zhou

Implementing the Multimodel Generalized Beta Estimator in Stata and Its Application

by

Yutong Duan, B.A.

Report

Presented to the Faculty of the Graduate School
of the University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science in Statistics

The University of Texas at Austin
May 2016

Abstract

Implementing the Multimodel Generalized Beta Estimator in Stata and Its Application

by

Yutong Duan, M.S.STAT.
The University of Texas at Austin, 2016

SUPERVISOR: Paul T. von Hippel
CO-SUPERVISOR: Mingyuan Zhou

The *multimodel generalized beta estimator*(MGBE) described by von Hippel, Scarpino and Hola (2014) provides researchers with an improved way to estimate inequality from binned incomes. To extend the application of MGBE, the *mgbe* command is developed in Stata. In this report, the implementation and performance of *mgbe* are discussed.

Table of Contents

Introduction

Studies in the social sciences are often interested in estimating household income inequality. When the data are in the form of binned or grouped data instead of the income of individuals, researchers may encounter a challenge in estimating a series of statistics such as mean, median, Gini coefficient, Theil index, and etc. from binned data. Sometimes the study only focuses on a specific area where there are a small number of bins, so researchers will want to find the distribution that fits the bins and acquire inequality statistics from the fitted distribution.

Table 1.1 shows the binned households' income of Autauga County, Alabama. Each county in the dataset has 16 bins, and the upper bound of the top bin is missing, because there is no upper limit of household income. The bottom bin technically also has no lower bound, but we treat the lower bound as 0.

Although numerous previous studies have discussed estimating inequality statistics from binned data, the estimation of binned data is not well supported in Stata, which is one of the most widely used software packages by the economists. Motivated by this, a new *mgbe* command has been written in Stata which implements a multimodel generalized beta estimator (MGBE) to estimate inequality statistics from binned data.

The *mgbe* command can fit the binned data into a single distribution selected from 10 available distributions by the user. The program will run over each group

and return estimates from the binned data. In addition, since the true distribution is unknown, the *mgbe* also supports model selection according to Akaike Information Criterion (AIC) or Bayes Information Criterion (BIC), or obtain estimates averaged across distributions.

The implementation of MGBE is available in R as a *binequality* package (Scarpino, von Hippel, and Holas 2014), which provides us with a good reference by which to compare its performance. In this report, the implementation of *mgbe* in Stata will be discussed, as will the performance of *mgbe* and compared with that of the *binequality* package.

Bin	**Households**	**Minimum**	**Maximum**
1	165	$0	$9,999
2	125	$10000	$14,999
3	104	$15,000	$19,999
4	111	$20,000	$24,999
5	150	$25,000	$29,999
6	109	$30,000	$34,999
7	125	$35,000	$39,999
8	139	$40,000	$44,999
9	118	$45,000	$49,999
10	241	$50,000	$59,999
11	275	$60,000	$74,999
12	368	$75,000	$99,999
13	202	$100,000	$124,999
14	118	$125,000	$149,999
15	79	$150,000	$199,999
16	38	$200,000	

Note: The income is adjusted to 2010 US dollars

Table 1: Distribution of Household Income in Autauga County, Alabama

Implementation

2.1 Methods

Maximum likelihood can be used to fit continuous densities to binned data (McDonald & Ransom 2008) [3]. The likelihood is calculated as

$$L = \prod_{b}^{B} (F(M_b) - F(m_b))^{n_b}$$

where M_b and m_b are the upper and lower bounds of bin b (von Hippel 2015) [8]. The log-likelihood, which takes log of L is written as

$$l = \sum_{b}^{B} n_b (F(M_b) - F(m_b))$$

which will be maximized to find the parameter of the fitted distribution. Since some inequality statistics cannot be directly calculated from GB parameters, we draw 1,000 evenly spaced quantiles from fitted distribution to calculate the median and inequality statistics. This alternative numerical approach is relatively accurate to estimate statistics from fitted distributions (McDonald and Ransom 2008) [3].

Among all 10 distributions from the GB family, the generalized beta type 2, Singh-Maddala and Dagum distributions have been implemented in Stata by Nichols, where users can fit binned data with *gbgfit* [6], *dagfit* [5] and *smgfit* [7] separately. They maximize based on an alternative definition of likelihood L_{alt} and log-likelihood l_{alt}. However, these three commands do not correctly calculate the log-likelihood

for the maximization process. The log-likelihood returned from these four commands are positive, which should always be negative according to the definition of log-likelihood. As von Hippel (2015) [8] points out:

> "Some publications give an alternative definition of the likelihood and log likelihood which includes factorial constants (Bandourian, McDonald,& Turley, 2002; McDonald, 1984; McDonald & Ransom, 1979; McDonald & Xu, 1995):
>
> $$L_{alt} = L * N! \prod_{b}^{B} \frac{1}{n_b!}$$
>
> $$l_{alt} = l + ln(N!) - \sum_{b=1}^{B} ln(n_b)$$
>
> The factorial constants are not incorrect, but they are unnecessary. The maximum likelihood estimates will be the same whether the factorial constants are included or not."

The *mgbe* command corrects the log-likelihood maximization part and combines all 10 distributions from GB family together. In addition, since the true distribution of each group is unknown, *mgbe* enables model selection among user-selected distributions according to either Akaike Information Criterion (AIC) or Bayes Information Criterion (BIC), or obtain estimates from averaging selected GB family distributions. The AIC is defined as

$$AIC = -2 * l + 2 * k$$

where l is the maximized log-likelihood and k is the number of parameters estimated. And BIC is defined as

$$BIC = -2 * l + k * lnN$$

where N is the sample size.

For our mean and variance, the formulas of mean and variance of each distribution are clearly defined and easy to calculate in Stata, so we use GB parameters to calculate the mean and variance. However, for the median and inequality statistics, the formulas sometimes may be unknown or the related calculations are not supported in Stata. Thus, we approximate these statistics by calculating from 999 quantiles of the fitted distribution. The estimate calculated in this way is only an approximate, but we found it has high reliability and high correlation to estimates derived from distribution parameters.

The 10 distributions implemented in the *mgbe* command starts from 4-parameter GB2 distribution, and the other distributions can be expressed as a special case of distribution or expressed as limiting certain parameters to infinity. Figure 1.1 shows part of the GB family relationship that we use in *mgbe*, where in addition to the 4-parameter GB2 distribution we have four 3-parameter distributions—Singh-Maddala, Dagum, beta 2 and generalized gamma distribution, and five 2-parameter distributions—log-logistic, Pareto 2, gamma, Weibull and lognormal distribution.

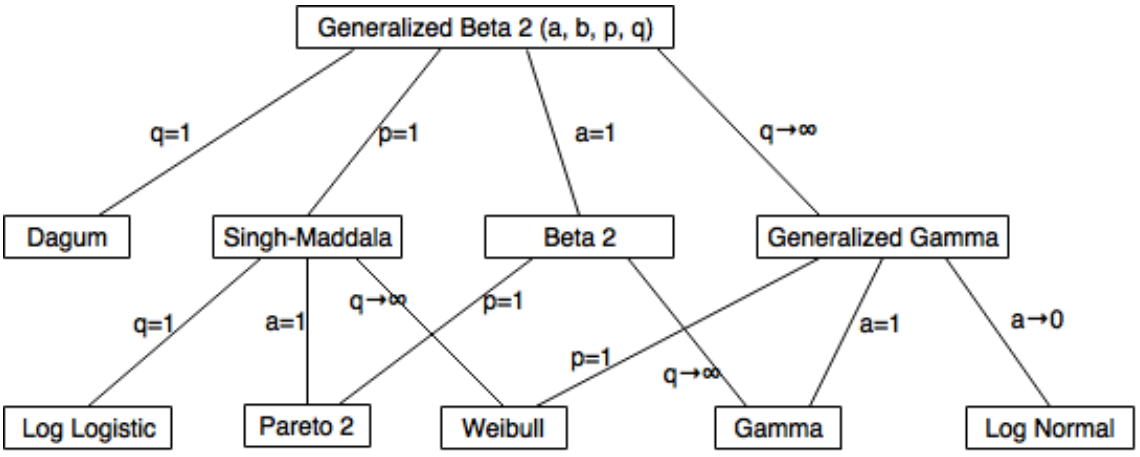


Figure 1: GB2 family distribution tree used in *mgbe*

For our *mgbe* command, the Singh-Maddala, Dagum, beta 2 and log-logistics distributions are nested in GB2 distribution by imposing constraints on parameters according to the GB family tree in Figure 1.1. For these cases, the cumulative density function(CDF) of the GB distribution is defined as [3] :

$$F(x;a,b,p,q) = I\left(\frac{(x/b)^a}{(1+(x/b)^a)}, p, q\right) \tag{2.1}$$

where $I(x,p,q)$ is the incomplete beta function (cumulative beta distribution).

And the quantile function which derives from the inverse of CDF by finding the value of Q for which $y = Q(u) = F^{-1}(u)$, where u is the percent of $Q(u)$ and $Q(u)$ is namely the percentile of u. The quantile function of the GB2 distribution derived from (2.1) is:

$$Q(u;a,b,p,q) = b * (IB(p,q,u)/(1-IB(p,q,u))^{1/a} \tag{2.2}$$

where $IB(.)$ is the inverse cumulative beta function.

The generalized gamma distribution is a special case of GB2 distribution where q is limited to infinity, so the generalized gamma distribution and distributions nested in generalized gamma—gamma and Weibull distributions have CDF [4]

$$F(x;a,b,p) = \Gamma\left((\frac{x}{b})^a, p\right) \tag{2.3}$$

where $\Gamma(x,p)$ is the incomplete gamma distribution (cumulative gamma distribution).

With the quantile function

$$Q(u;a,b,p) = IG(p,u)^{1/a} * b \tag{2.4}$$

where IG is the inverse of the incomplete gamma function.

The lognormal distribution is also defined separately since it is a special case of generalized gamma distribution where a is limited to 0. In this case, the lognormal distribution has to be defined separately from generalized gamma distribution. Here the CDF of lognormal distribution is defined as

$$F(x; b, p) = \Phi\left(\frac{ln(x) - b}{p}\right), x > 0 \tag{2.5}$$

where Φ is cumulative standard normal distribution

And the quantile function of lognormal distribution is

$$Q(u; b, p) = exp(b + p * IN(u)) \tag{2.6}$$

where IN is the invert of Normal distribution.

The Pareto 2 distribution can be expressed by constraining $a = 1$ and $p = 1$ on GB2 distribution. However, in this program simply imposing constraints does not work well to get parameter estimates, so Pareto 2 is defined separately from GB2 distribution with CDF [1]

$$F(x; b, q) = 1 - \left(\frac{b}{x + b}\right)^q \tag{2.7}$$

And its quantile function is

$$Q(u; b, q) = b * ((1 - u)^{-1/q} - 1) \tag{2.8}$$

2.2 Implementation in Stata

The program consists of the main program *mgbe.ado*, along with two other subprograms *mgbe_ll.ado* and *mgbe_cdf.ado*. The *mgbe_cdf.ado* created as a separate file is to make the program neat and easier to maintain, all the *CDF*s of the available

distributions are in *mgbe_cdf.ado* and the log-likelihood is defined in *mgbe_ll.ado* when related distributions are selected in the main *mgbe.ado* and the log-likelihood is maximized using the *ml model* maximization command.

In the *mgbe_ll.ado* subprogram, the function of CDF is called from *mgbe_cdf.ado* where the distribution is identified and the corresponding CDF will be calculated. When income is 0, CDF will return 0, and when income is infinite as it is for the right side of the top bin, CDF will return 1. For other values of income, the value of CDF depends on the distribution being fitted (see section 2.1 for the CDFs of the available distributions). In the main program *mgbe.ado*, Stata will maximize the log-likelihood with *ml model*, which calls the *mgbe_ll.ado* subprogram to calculate the log-likelihood for each selected distribution and each group. To increase the maximization speed here, parameter values are initialized to values typical for fitted distributions or a maximization *technique* is used or both methods are adopted to increase the speed for some distributions that take many iterations to converge. Notice that if the number of populated bins is smaller than the parameter plus 1, the distribution is unidentified, the maximization process will not run and all estimates will return as missing values.

The outputting of *mgbe* captures the model parameters, mean, variance, median, inequality statistics such as Gini coefficient, Theil index, coefficient of variation (CV) etc. as well as 999 quantiles. Before outputting the results, *mgbe* will check if the moments are defined. When the mean is infinite or undefined (e.g. $1/a > q$ or $a/1 < -p$ for GB2 and those distributions nested in GB2 distributions; $1/a < -p$ for generalized gamma and distributions nested in generalized gamma distribution; $q < 1$ for Pareto 2 distribution), the mean and inequality statistics are reported as missing values in the Stata outputting. And when the variance is infinite or

undefined (e.g. $2/a > q$ or $a/2 < -p$ for GB2 and those distributions nested in GB2 distributions; $2/a < -p$ for generalized gamma and distributions nested in generalized gamma distribution; $q < 2$ for Pareto 2 distribution), in addition to reporting missing values for mean and inequality statistics, the variance is also reported as missing.

When implementing the *mgbe* in Stata, we found the following issues with their solutions or possible solutions

1. The maximization of the log-likelihood will only converge if calculated precisely using double precision floating point numbers.

2. The maximization *technique*, which is an option in *ml model*, is coded as default for each distribution in this program, that is, users cannot select the -technique- by themselves. However, there is no a universal maximization option that fits all the 10 distributions. So each maximization process is separately specified by testing on a lot of combinations of options. We found that the *technique(dfp 5 nr 5)*, where *dfp* is the Davidon-Fletcher-Powell formula in optimization and *nr* is the Newton-Raphson method in optimization, is the best maximization technique for most distributions which require more iterations to achieve their maximized log-likelihood values. If the step is too big (i.e., dfp 20 or nr 20), the program will produce a lot of non-convergent cases. While if the step is too small (i.e. dfp 2 nr 2), the frequent interchange of the two methods will slow down the process to find the maximized log-likelihood value.

3. Maximizing the log-likelihood of each group takes time. Although it seems acceptable that it takes around four seconds to run 10 distributions on each county, this still makes *mgbe* difficult to apply to very large datasets with 3,221 counties. When the dataset is large, the steps after maximizing log-likelihood such as reshaping

the dataset and generating inequality statistics will also take longer time to process. In addition to the maximization of log-likelihood, calculating 11 inequality statistics with *egen_inequal* also takes a longer time than we want. A number of steps were taken to reduce runtime, and users wishing to reduce runtime further can select, for example, 3 of the 10 distributions.

2.3 Software

What motivates this project is that although the *binequality* package is available for MGBE estimation in R (Scarpino, von Hippel, and Holas 2014), we want to compare results by running the MGBE estimator in Stata as an alternative implementation. Compareing two implementations will expose problems, and Stata is more widely used by economists.

Since *mgbe* will generate 999 quantiles for each group, *mgbe* has to run on versions at least as large as Stata/IC.

Data

The *mgbe* command starts by running on households income data from 3,221 counties in the United States and Puerto Rico, collected from the American Community Survey (ACS). ACS pooled data from 2006-2010 and sampled 1 in 8 to increase accuracy. All the income data are adjusted to 2010 US dollars. There are 16 bins in each county, see Table 1.1 for an example.

To evaluate the performance of *mgbe*, the estimates from *mgbe* are compared with the true mean, median and Gini of each county, which are published by the Census (US Census Bureau). The true values that we use here are calculated by Census Bureau from unbinned incomes. Since the Census does not have income data for every household, the true Gini of population is unknown, but the data from the Census should be more accurate than our estimates from binned data, so it will provide us with a good reference. In addition to comparing our results to the true values, we also compare the output of our Stata implementation with that from the same method *binequality* implemented in R.

Results

4.1 GB estimates

In this report, we are focusing on the accuracy of mean, median and Gini estimates, and comparing the output with output obtained by von Hippel, Scarpino and Holas[9] by running R's *binequality* package.

Table 4.1 shows the bias, RMSE (percent root mean squared error) and reliability of median, mean and the Gini estimate. The *bias* is defined as the mean of estimate minus the true value. The *RMSE* is evaluated by the square root of the mean squared error. And the *reliability* is the squared correlation between the estimate and true value. Both the bias and the RMSE are expressed as percentages of the true parameter value.

The results from the *mgbe* command are generally consistent with the results from R's *binequality* package for most distributions. The most inconsistent cases lie in Pareto 2 and beta 2 distributions, as well as the reliability of Gini estimates from the Singh-Maddala distribution. For the Pareto 2 distribution, the reliabilities of median and mean from the *mgbe* are higher than those of *binequality*'s output, while the reliability of the Gini estimate is lower than that of R output. The bias is in the same direction, and the RMSEs of estimates from *mgbe* are generally 3-4% lower than the RMSEs from *binequality*. The instability of accuracy may result

from the poor fit of Pareto 2 distribution, as shown in Figure 4.2, which graphs the mean, median and Gini estimates from *mgbe* and true estimate values of Pareto 2 distribution. Since the Gini estimate is calculated as $q/(2q-1)$ and q is constrained greater than 1, the Gini estimates are always greater than 0.5, which also implies why the Pareto 2 is not a good fit for estimating our households' income Gini. Figure 4.1 also gives the performance of GB2 distribution, which is one of the best models from Table 4.1 as well as its probability being selected by AIC or BIC in Table 4.3. If the distribution is a good fit to the binned data, the estimates from fitted distribution should scatter closely around the true values, as we see in Figure 4.1, compared to the poor fitting from the Pareto 2 distribution in Figure 4.2. In addition, the negative bias of the mean, median and Gini estimates fitted from GB2 can also be seen from Figure 4.1.

Beta 2 distribution generally has higher reliability for mean, median and Gini estimate, especially as the reliability of Gini estimate is 11% higher than that from R output. However, the mean is 2% negatively biased compared to no bias reported from R's output, and the Gini estimate is 2% negatively biased while the Gini estimate from R output is 1 % positively biased.

	Median						Mean						Gini					
	% Bias		% RMSE		%Reliable		% Bias		% RMSE		%Reliable		% Bias		% RMSE		%Reliable	
	Stata	R	Stata	R	Stata	R	Stata	R	Stata	R	Stata	R	Stata	R	Stata	R	Stata	R
Weibull	2	2	4	4	98	98	-2	-2	4	4	98	98	-4	-4	5	5	83	83
Log Logistic	-5	-5	7	6	97	99	14	12	15	13	95	97	13	13	14	14	46	47
Pareto 2	-12	-16	14	17	95	91	-2	-1	3	6	98	93	19	20	16	22	21	23
Gamma	0	0	4	4	98	98	-2	-1	4	4	98	98	-4	-4	5	5	80	81
Log Normal	-9	-9	9	9	98	98	3	3	5	5	98	98	6	6	7	8	72	73
Dagum	1	1	4	3	99	99	1	1	3	3	99	99	-1	0	4	4	86	83
SM	0	-2	4	4	99	97	-2	-1	3	3	99	97	-3	-1	5	6	86	77
Beta 2	-3	-5	4	6	99	98	-2	0	3	4	99	98	-2	1	4	5	84	73
GG	-1	-1	3	3	99	99	-2	-2	3	3	99	99	-3	-4	5	5	86	86
GB2	-1	0	3	3	99	98	-2	-1	3	3	98	97	-3	-1	4	3	86	87

Table 2: Comparing Median, Mean and Gini Estimates with R Output

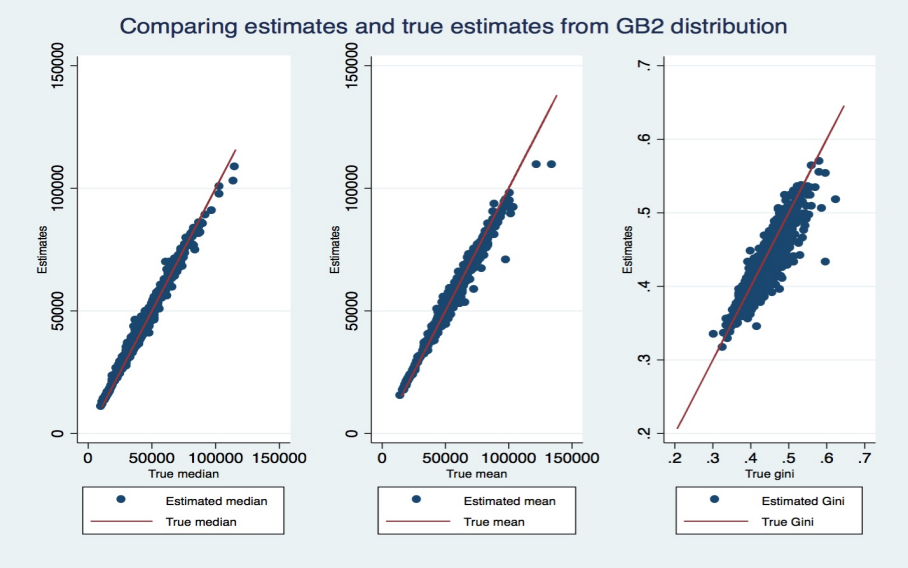


Figure 2: Comparing estimated and true median, mean and Gini estimates from GB2 Distribution

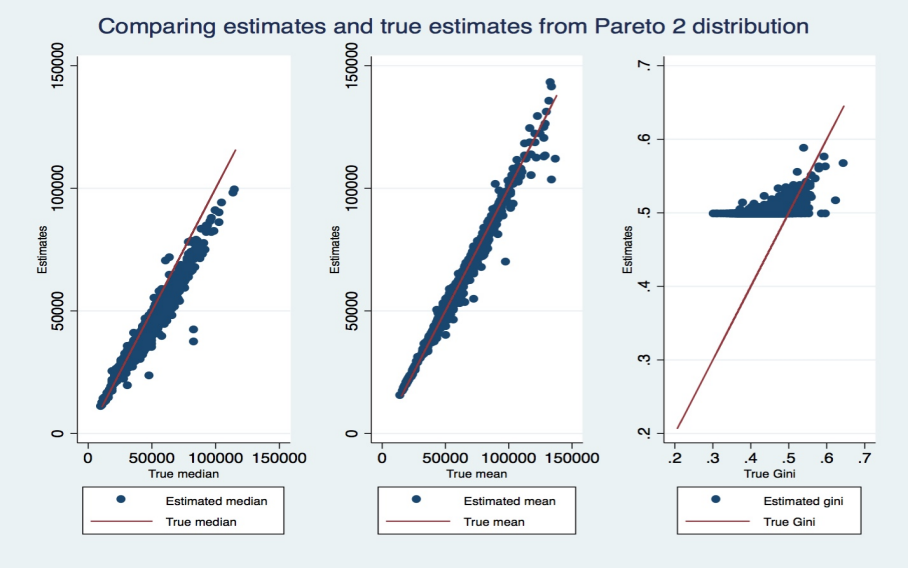


Figure 3: Comparing estimated and true median, mean and Gini estimates from Pareto 2 Distribution

Lastly, the Singh-Maddala distribution in *mgbe* command gives a slightly higher reliability of median and mean and 9% higher reliability of Gini estimate. In addition, the estimated mean from *mgbe* command has no bias, while the reliability of mean from R output is negatively 2% biased from true median values.

In addition to the estimates accuracy, the average iterations that each distribution to achieve convergence are also recorded in Table 4.2. The maximum number of iteration is set as 200 (default is 16,000 in Stata), the maximization technique is *dfp 5 nr 5* and the tolerance for the scaled gradient is 1e-3. If the maximization fail to finish within 200 iterations, the program will show a warning. In addition, if the maximization tries too many iterations before convergence for many groups, it may indicate that changing maximization parameters or parametrization method is required.

Distribution	**Average Iterations**
Weibull	4
Log Logistic	7
Pareto 2	11
Gamma	5
Log Normal	5
Dagum	5
Singh-Maddala	9
Beta 2	13
GG	11
GB2	18

Table 3: Average Iterations to Achieve Convergence

4.2 Multimodel selection

The *mgbe* command also implements model selection by Akaike Information Criterion (AIC) or Bayes Information Criterion (BIC) or returning a weighted average model with averaged estimates. Table 4.3 shows the percentage of each distribution selected under each selection criteria. Selecting via AIC or BIC does not always leads to exactly the same best-fitting model, which is especially obvious in the case of GB2 distribution. And the percentages of some other distributions being selected in the *mgbe* output are also slightly different from that of the R output.

The GB2 distributions account for more than half of the best-fitting model under AIC and more than one-third under BIC running from *mgbe*, while only more than 10% that GB2 is selected as best-fitting model running from *binequality* in R. There's also an increase of percentage that the Beta 2 distribution and Singh-Maddala distribution perform as a best-fitting scenario. The increase in these cases is mostly offset by the decrease of the percentage that Dagum distribution and gamma distribution are selected as best-fitting model compare to the results from R.

By weighting distributions selected by users by AIC or BIC, *mgbe* can obtain averaged estimates of each county according to its weights. The weighted estimates can be written as

$$\hat{\bar{\theta}} = \sum_{i=1}^{R} w_i \hat{\theta}_i$$

where we are weighting over models $i = 1, 2, 3, ..., R$[2]. Table 4.3 gives the weights that each distribution accounts for the final averaged model, which performs in a very similar way discussed for AIC and BIC selection.

The estimates accuracy is recorded in Table 4.4, where it shows that the reliabilities

	%Selected				Weighted			
	by AIC		by BIC		by AIC		by BIC	
	Stata	R	Stata	R	Stata	R	Stata	R
Weibull	2	6	3	7	3	6	5	7
Log Logistic	0	1	0	1	1	1	1	2
Pareto 2	0	0	0	0	0	0	1	0
Gamma	3	21	7	24	3	14	7	17
Log Normal	1	3	1	3	3	2	4	3
Dagum	11	28	19	29	11	23	18	24
Singh-Maddala	9	7	13	8	8	10	12	10
Beta 2	5	1	7	1	5	3	7	3
GG	15	19	13	16	16	21	14	19
GB2	53	15	36	11	50	21	31	17

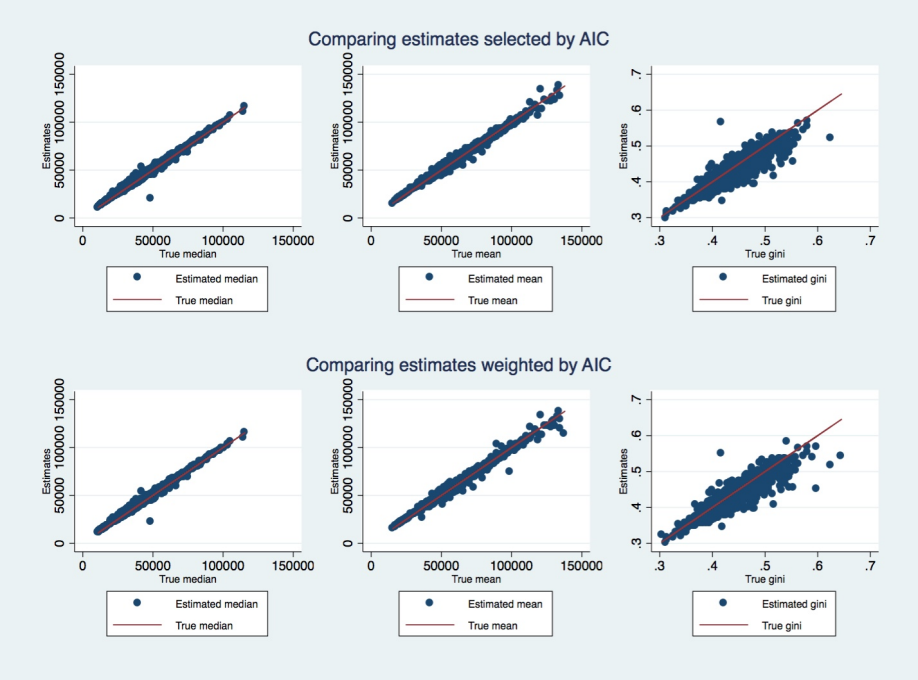
Table 4: Multimodel Selection

and RMSEs perform at least as well as those from any single distribution. While the bias of mean and Gini estimates cannot outperform Dagum or beta 2 distribution by themselves, the difference is only around 1% for both cases though. Moreover, although the probability that the same distribution selected as the best-fitting model varies a lot under different model selection criteria as shown in Table 4.3, the bias, RMSE and reliability are stable for all these four methods. The scatterplot in Figure 4.3 gives the performance under each selection criteria, where we may see that the *mgbe* typically gives accurate estimates.

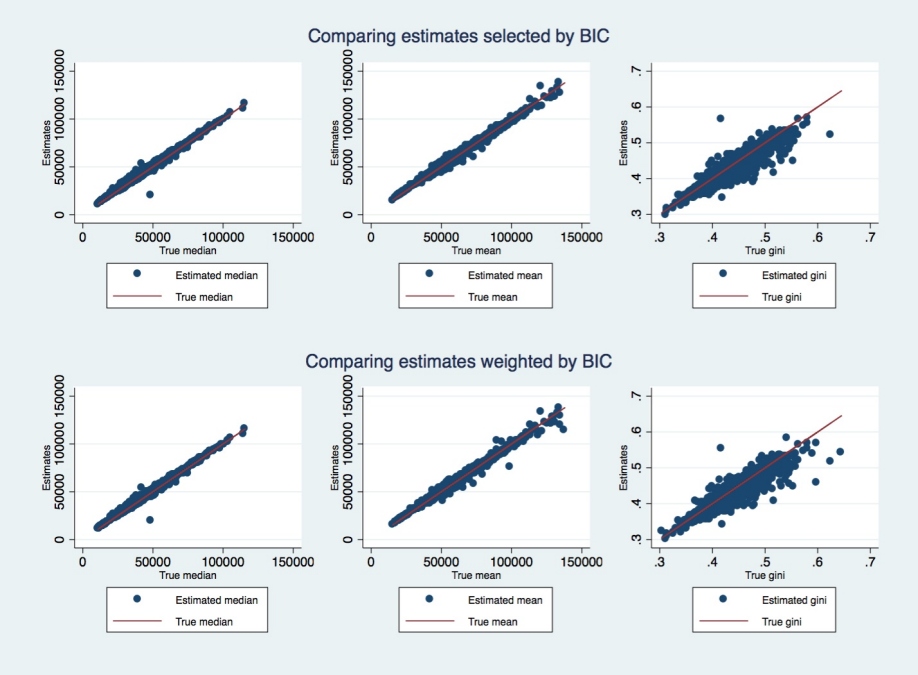
What's similar for the results from both Stata and R is that both of the Pareto 2 and log-logistic distributions are almost never selected, and the lognormal distribution also has a small chance to be selected. This may result from a large percentage of undefined moments for log-logistic distribution and a poor fitting of Pareto 2 and lognormal distribution.

	Median			Mean			Gini		
	% Bias	% RMSE	%Reliable	% Bias	% RMSE	%Reliable	% Bias	% RMSE	%Reliable
AIC	0	3	99	-1	3	99	-3	4	88
BIC	0	3	99	-1	3	99	-3	4	87
Weighted by AIC	0	3	99	-1	3	99	-3	4	87
Weighted by BIC	0	3	99	-1	3	99	-3	4	87

Table 5: Estimates Bias, RMSE and Reliability from Weighted Average Models



(a) AIC



(b) BIC

Figure 4: Compare estimates with their true values by multimodel selection

Discussion

One thing that may need to improve for *mgbe* command is about the running speed, and that is we are always trying to optimize. The total running time over the 10 distributions on 3,221 counties is 5.4 hours. It is tested with Stata/MP 14.1 on a Mac with 3.5 GHz i7 processor and 32GB memory. The program slows down as the size of dataset increases, so it would be more applicable on large datasets if we can make the program run as fast as possible.

In addition, *mgbe* does not work on groups with few bins. For example, Loving County, Texasm only has two populated bins, so it cannot be fitted by any distributions with more than one parameter. Further studies may examine how *mgbe* works on these cases where the number of bins is small. Moreover, when testing on the *mgbe* command, we also notice that different maximization options in Stata will also affect the results of estimates, which may have an effect on the returned estimates when the number of bins is very small.

Conclusion

The *mgbe* command in Stata provides researchers with an integrated method to fit GB2 family distributions on binned income data as an alternative tool for the *binequality* package in R. Users can choose a series of distributions from GB family as well as the multimodel method to choose the best-fitting model. The accuracy that *mgbe* performs on binned data is high for most cases and it produces similar results with R's *binequality* package as introduced in a previous study. While an advantage of the *binequality* package is the running speed. But this new command enables users to implement a multimodel GB estimator in Stata with more extended applications and better accuracy than currently available GB estimator commands in Stata.

Bibliography

[1] David E. Giles, Hui Feng, and Ryan T. Godwin. On the bias of the maximum likelihood estimator for the two-parameter lomax distribution. *Communications in Statistics - Theory and Methods*, 42(11):1934–1950, 2013.

[2] William Gould, Jeffrey Pitblado, William Sribney, and Stata Corporation. *Maximum likelihood estimation with stata*. Stata Press, College Station, Tex, 3rd edition, 2006.

[3] James. McDonald and Michael. Ransom. *Modeling Income Distributions and Lorenz Curves*, chapter The Generalized Beta Distribution as a Model for the Distribution of Income: Estimation of Related Measures of Inequality, pages 147–166. Springer New York, New York, NY, 2008.

[4] William Q. Meeker, Luis A. Escobar, and Inc NetLibrary. *Statistical methods for reliability data*. Wiley, New York, 1 edition, 1998.

[5] Austin Nichols. *DAGFIT: Stata module to fit a Generalized Beta (Type 2) distribution to grouped data via ML*, 2010.

[6] Austin Nichols. *GBGFIT: Stata module to fit a Generalized Beta (Type 2) distribution to grouped data via ML*, 2010.

[7] Austin Nichols. *SMGFIT: Stata module to fit a Singh-Maddala distribution to grouped data via ML*, 2010.

[8] Paul T. von Hippel. The likelihood for binned data. November 2015.

[9] Paul T. von Hippel, Samuel V. Scarpinom, and Igor Holas. Robust estimation of inequality from binned incomes. *ArXiv e-prints*, February 2014.